The second eigenvectors of the Google matrix and their relation to link spamming

SIAM Student Chapter

Alex Sangers
joint work with Martin van Gijzen

Delft University of Technology

September 19, 2014

Outline

Scientific fame and glory

What determines scientific fame?

- Number of publications?
- Number of citations?
- h-index?
- Erdös number?

Scientific fame and glory

What determines scientific fame?

- Number of publications?
- Number of citations?
- h-index?
- Erdös number?

It is your PageRank.

PageRank and link spamming

- PageRank: the importance of a web page
- Link spamming: increase your PageRank

PageRank and link spamming

- PageRank: the importance of a web page
- Link spamming: increase your PageRank

The research:

- What is the relation between link spamming and the second eigenvectors of the Google matrix?
- Develop an algorithm to compute a complete set of independent eigenvectors for the second eigenvalue.

A model of the web surfer

Web pages are connected through hyperlinks.

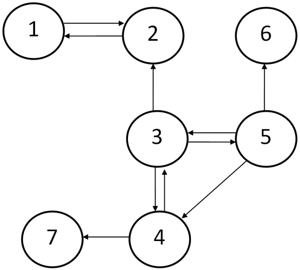


Figure: Model of part of the web

A model of the web surfer

Web pages are connected through hyperlinks.

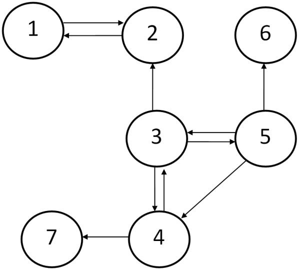


Figure: Model of part of the web

Mathematical model:

- Binary matrix $\mathbf{G}$, with $\mathbf{G}_{i,j} = 1$ if page j links to page i.
- Row-stochastic transition matrix $\mathbf{P}$.

The matrices G and P

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

The matrices G and P

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{P}^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 1/7 & 1/7 \\ 1 & 0 & 1/3 & 0 & 0 & 1/7 & 1/7 \\ 0 & 0 & 0 & 1/2 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 1/3 & 0 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 1/3 & 0 & 0 & 1/7 & 1/7 \\ 0 & 0 & 0 & 0 & 1/3 & 1/7 & 1/7 \\ 0 & 0 & 0 & 1/2 & 0 & 1/7 & 1/7 \end{pmatrix}$$

Teleportation

Teleportation is jumping to a web page without following a link.

$$\mathbf{A} = p\mathbf{P}^T + \frac{1-p}{N}\mathbf{e}\mathbf{e}^T$$

Here, $\mathbf{e} = [1, 1, \ldots, 1]^T$.

Teleportation

Teleportation is jumping to a web page without following a link.

$$\mathbf{A} = p\mathbf{P}^T + \frac{1-p}{N}\mathbf{e}\mathbf{e}^T$$

Here, $\mathbf{e} = [1, 1, \ldots, 1]^T$.

Probability p:	Follow an outlink,
Probability $1-p$:	Teleport to any web page.

The PageRank

- PageRank vector $\mathbf{x}$ is the probability vector after infinitely long surfing.
- Google matrix $\mathbf{A}$ has a unique largest eigenvalue $\lambda_1 = 1$ (by Perron-Frobenius).
- PageRank vector $\mathbf{x} > 0$ is the eigenvector of $\mathbf{A}$ corresponding to $\lambda_1 = 1$.
- Compute the PageRank vector by the Power method.

Link spamming

How can you increase your PageRank by using link structures?

Irreducible closed subchains

An irreducible closed subchain:

- Once entered a closed subchain, you cannot leave;
- Every node in an irreducible subchain can be reached.
- Nodes in an irreducible closed subchain receive a high PageRank value.

Irreducible closed subchains

An irreducible closed subchain:

- Once entered a closed subchain, you cannot leave;
- Every node in an irreducible subchain can be reached.
- Nodes in an irreducible closed subchain receive a high PageRank value.

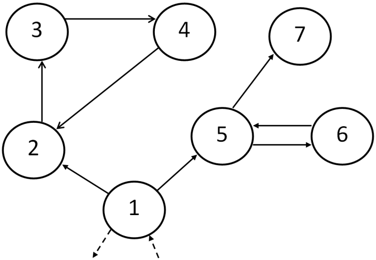
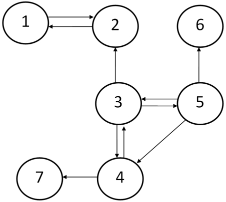
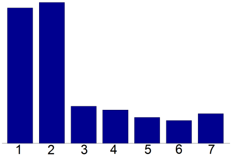


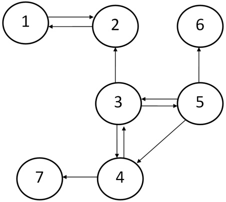
Figure: Irreducible closed subchains?

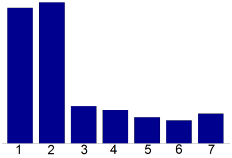
Link spamming: node 4



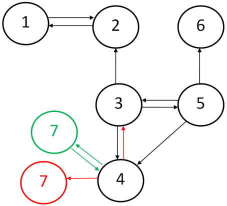
with $\mathbf{x}^{(1)} =$ 

Link spamming: node 4

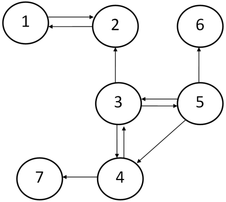


with $\mathbf{x}^{(1)} =$ 

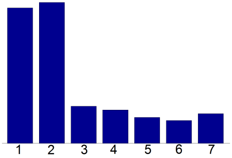
$\Downarrow$



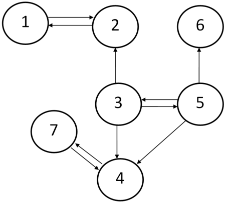
Link spamming: node 4



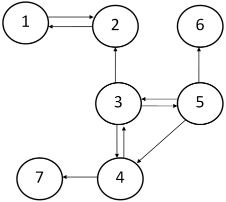
with $\mathbf{x}^{(1)} =$



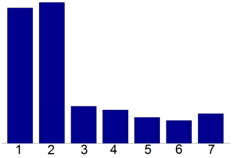
$$\Downarrow$$



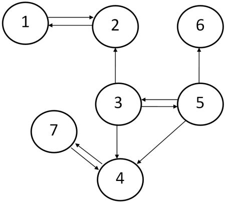
Link spamming: node 4



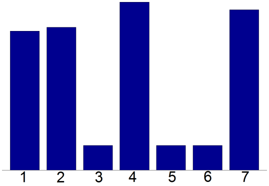
with $\mathbf{x}^{(1)} =$



$\Downarrow$



with $\mathbf{x}^{(1)} =$



Tarjan's algorithm

Detection of link spamming:

- Find the irreducible subchains of the graph.
- An irreducible subchain consisting of a group of nodes, without outlinks, is an irreducible closed subchain.

The second eigenvectors of $\mathbf{A}$

$$\mathbf{A}\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})},$$

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}.$$

The second eigenvectors of $\mathbf{A}$

$$\mathbf{A}\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})},$$

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}.$$

Note that

$$\mathbf{e}^T\mathbf{x}^{(\mathbf{2})} = 0,$$

by bi-orthogonality of the left eigenvector $\mathbf{e}^T$ and the right eigenvector $\mathbf{x}^{(\mathbf{2})}$ of $\mathbf{A}$.

The second eigenvectors of $\mathbf{A}$

$$\mathbf{A}\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})},$$

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}.$$

Note that

$$\mathbf{e}^T \mathbf{x}^{(\mathbf{2})} = 0,$$

by bi-orthogonality of the left eigenvector $\mathbf{e}^T$ and the right eigenvector $\mathbf{x}^{(\mathbf{2})}$ of $\mathbf{A}$.

$$\Rightarrow p\mathbf{P}^T \mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}$$

If $\mathbf{P}^T$ contains at least two irreducible closed subchains, then $\lambda = p$.

The second eigenvectors of $\mathbf{A}$

$$\mathbf{A}\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})},$$

$$(p\mathbf{P}^T + \frac{(1-p)}{N}\mathbf{e}\mathbf{e}^T)\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}.$$

Note that

$$\mathbf{e}^T\mathbf{x}^{(\mathbf{2})} = 0,$$

by bi-orthogonality of the left eigenvector $\mathbf{e}^T$ and the right eigenvector $\mathbf{x}^{(\mathbf{2})}$ of $\mathbf{A}$.

$$\Rightarrow p\mathbf{P}^T\mathbf{x}^{(\mathbf{2})} = \lambda \mathbf{x}^{(\mathbf{2})}$$

If $\mathbf{P}^T$ contains at least two irreducible closed subchains, then $\lambda = p$.

$$\Rightarrow \mathbf{P}^T\mathbf{x}^{(\mathbf{2})} = \mathbf{x}^{(\mathbf{2})}$$

Computation of all the second eigenvectors

To solve $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, we can solve

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

by applying IDR(s) with a nonzero starting vector.

Computation of all the second eigenvectors

To solve $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, we can solve

$$(\mathbf{I} - \mathbf{P}^T)\mathbf{x} = 0 \quad .$$

by applying IDR(s) with a nonzero starting vector.

The nonzero-elements in $\mathbf{x}$ correspond to nodes in an irreducible closed subchain. Thereafter, apply Tarjan's algorithm to find the *different* irreducible closed subchains.

Detection of link spamming

Two link spamming detection algorithms:

1. Search the web for irreducible closed subchains (Tarjan's algorithm)
2. - Compute a first eigenvector of $\mathbf{P^T}$,
   - Determine the nonzero entries,
   - Use Tarjan's algorithm only on the nodes corresponding to nonzero entries.

Results

Test problem	Size	Closed subchains	CPU-time Tarjan	CPU-time Eigenvector
wb-cs-stanford	9914	113	0.3	1.4
flickr	820878	5394	399.3	160.8
wikipedia-20051105	1634989	68	1515.3	140.2
wikipedia-20060925	2983494	63	5077.1	166.6
wikipedia-20061104	3148440	59	5696.9	155.1
wikipedia-20070206	3566907	58	7462.7	313.6
wb-edu	9845725	49573	75703.2	2825.6*

Computing time for web crawls by Gleich

Note: For `wb-edu` the eigenvector algorithm found 41606 subchains.

Conclusions

The relation between link spamming and the second eigenvector of the Google matrix:

- The second eigenvectors are combinations of the dominant eigenvectors of the matrix $\mathbf{P}^T$
- The dominant eigenvectors of $\mathbf{P}^T$ correspond to irreducible closed subchains
- Nodes in these subchains get artificially high PageRanks
- We have proposed an efficient algorithm to compute all second eigenvectors of the Google matrix and to detect this kind of link spamming